International Journal of
Geo-Information

MDPI

Article

Towards Detecting the Crowd Involved in Social Events

Wei Huang [1,*], Hongchao Fan [2] and Alexander Zipf [1]

[1] Institute of Geography, Heidelberg University, 69120 Heidelberg, Germany; zipf@uni-heidelberg.de
[2] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; hongchao.fan@whu.edu.cn
* Correspondence: wei.huang@uni-heidelberg.de; Tel.: +49-6221-54-5525

Received: 9 September 2017; Accepted: 13 October 2017; Published: 16 October 2017

Abstract: Knowing how people interact with urban environments is fundamental for a variety of fields, ranging from transportation to social science. Despite the fact that human mobility patterns have been a major topic of study in recent years, a challenge to understand large-scale human behavior when a certain event occurs remains due to a lack of either relevant data or suitable approaches. Psychological crowd refers to a group of people who are usually located at different places and show different behaviors, but who are very sensitively driven to take the same act (gather together) by a certain event, which has been theoretically studied by social psychologists since the 19th century. This study aims to propose a computational approach using a machine learning method to model psychological crowds, contributing to the better understanding of human activity patterns under events. Psychological features and mental unity of the crowd are computed to detect the involved individuals. A national event happening across the USA in April, 2015 is analyzed using geotagged tweets as a case study to test our approach. The result shows that 81% of individuals in the crowd can be successfully detected. Through investigating the geospatial pattern of the involved users, not only can the event related users be identified but also those unobserved users before the event can be uncovered. The proposed approach can effectively represent the psychological feature and measure the mental unity of the psychological crowd, which sheds light on the study of large-scale psychological crowd and provides an innovative way to understanding human behavior under events.

Keywords: human activity; human behavior; psychological crowd; social event; Twitter

1. Introduction

Social events, especially those that can drive a massive influx of participants from different places to be gathering at certain place or places, such as protests and social conflicts, influence people's normal life and the way urban systems work to a certain extent. As such, it is essential to study how the involved individuals' spatial patterns would be during the events, and, more importantly, detect the unobserved but related individuals, i.e., such individuals that cannot (or are hard) to be directly linked to a corresponding event. For example, some individuals who participate in a protest on a street for minimum wage can be the persons who are just interested in certain related things or whose friends or relatives are involved in the protest rather than themselves being the low-income people. Those unobserved participants may boost some unexpected activities, such as criminal activities, which can become out of control.

The emerge of the related theoretical study on how a crowd could be shaped upon an event occurring can be traced back to the 19th century [1], where Ref. Le Bon [2] argued that a group of people who are usually located at different places and show different behaviors can be very sensitively driven to take the same act by a certain event. Such a crowd was defined as a psychological crowd.

A psychological crowd may consist of half a dozen individuals at certain moments, while it may not happen if hundreds of people gathering together by accident, such as thousands of people accidentally gathered in a public place without any determined object. In other words, the related individuals may only geographically gather together to show the collective behavior when a certain event occurs, such as a protest, during a short period. With regards to collective behavior, Turner et al. [3] argued that the mental unity of the crowd, i.e., the crowd behaves as one, needs to be explained.

Recently, there has been work on computationally understanding and modeling crowd behavior in certain environments or under certain events from small-scale. For example, Helbing et al. [4] studied pedestrian move flows in the scenario with a variety of facilities; Moussaïd et al. [5] proposed a cognitive science approach for the study of crowd disasters and the simulation of pedestrian flows under events; Sieben et al. [6] explored the social psychology of pedestrian dynamics during the assembling and dispersal of gatherings; Ref. von Krüchten and Schadschneider [7] performed an empirical study on investigating the impact of social groups on evacuations. However, a computational way to model certain psychological features of a crowd from a large-scale (e.g., national level) still lacks [8], which is also crucial for better understanding the way people live, work and play in cities—human activity patterns.

A comprehensive understanding of human activity patterns within urban environments (large-scale crowd behavior) lays a foundation to a variety of urban studies, ranging from transportation to social events [9–14]. Researchers have placed efforts on uncovering the urban spatial patterns of human activities. González et al. [15] found that a few of the locations intend to be frequently revisited, which can be characterized by a single spatial distribution indicating the dynamics behind the reproducible scaling patterns. Song et al. [16] reported that 93% of human movements can be potentially predicted based on an analysis of phone call data. Lima et al. [17] analyzed personal car GPS trajectories and found that most drivers use a small number of routes for their routine trips, and tend to have a preferred route for frequent trips. Jiang et al. [18] developed a framework for extracting mobility patterns from raw mobile phone data, where a pipeline that can translate the mobile phone records into interpretable spatial human mobility patterns was designed. However, such studies do not address the activity patterns of those unobserved people who are actually involved in an event. In fact, their behaviors when a certain event happens are rarely predictable. Inspired by the concept of psychological crowd, modeling psychological features of mass crowd can be a solution of detecting the unobserved individuals related to an event.

Massive multi-dimensional data that can represent human urban behaviors have been unprecedentedly collected in recent years, covering both spatiotemporal and semantic information (e.g., GPS trajectories; phone call data and social media data). These data have shown a power to infer human activities, sentiments, opinions and interests [19–21], which are also a basis of investigating the psychological features of the crowd. In this paper, we aim to propose a computational approach to detect the psychological crowd from a large-scale, where both observed and unobserved members can be detected. First, the psychological features of the crowd are modeled from a semantic manner using a machine learning method. Subsequently, semantic similarity is computed to detect the individuals who show mental unity. Based on the semantic similarity, a psychological crowd is finally derived, consisting of such individuals. A case study is conducted to test the proposed approach and analyze the characteristics of the detected crowd.

2. Methodology

2.1. Preliminaries

We assume that the psychological features can be depicted by the involved individuals' attributes. The involved individuals may and may not be geographically and temporally similar before a certain event occurs, but will show up together in a place or certain places during an event, as shown in Figure 1.

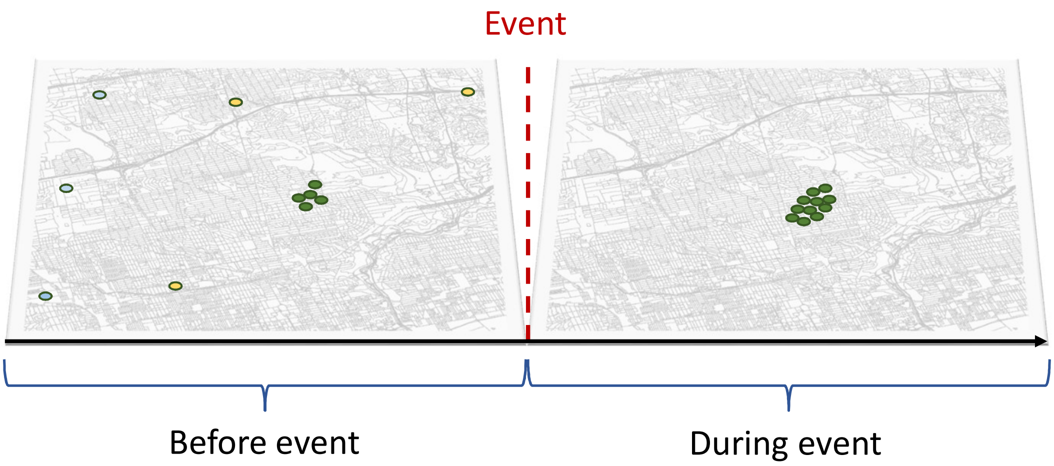


Figure 1. Geospatial-temporal distribution of a psychological crowd before and during an event. The circles refer to the involved individuals. The filled color of circles refers to the temporal pattern of the individuals.

Definition 1. *An* ***individual's attribute****, $P(a)$, is spatiotemporally irrelevant, which can reflect people's interests, sentiments, activities, opinions and attitudes on certain events, and activities participated in. The topics people talk about in daily life are used to infer the attribute.*

Definition 2. ***Mental unity*** *of a crowd indicates the involved individuals are similar in terms of the attribute. According to* ***Definition 1****, the similarity of the attributes can be measured by the similarity of the topics—semantic similarity. In this context, the individuals involved in a crowd should be the individuals who are semantically similar.*

2.2. Psychological Feature Modeling

Topic modeling, a machine learning method, is used to infer the topics representing the individual's attribute. Topic modeling has been effectively used to analyze textual information, inferring semantic patterns [22–25]. The idea of topic modeling is that, in a corpora, each document is a mixture of topics, where each topic is a mixture of words. The distribution of the topics in each document is used to represent the meaning of a document. The distribution of the words in each topic is used to represent the meaning of a topic. In our case, each people corresponds to the document and the distribution of topics corresponds to the individual's attribute—the psychological feature. Different techniques have been developed to estimate the distributions, among them, Latent Dirichlet Allocation (LDA) [22] is widely used. Thus, we use LDA in this study to estimate the distribution of topics and the distribution of words in each topic.

LAD is a generative model specifying a simple probabilistic procedure, in which any textual document can be drawn. First, a new document is created by choosing a distribution overt topics. Subsequently, for each word in that document, a topic is randomly chosen and a word is drawn from that topic according to a distribution over words. In order to infer the topic cluster involved in generating a collection of documents and the word cluster involved in generating a collection of topics, the generating process is inverted by using statistical techniques.

We assign the distribution over topics to $P(a)$ in Definition 1 and use $P(w|a)$ to represent the distribution over words in each topic. Only words (w) can be observed in the model. We use $\phi(n)$ to represent the multinomial distribution over words for topic n, i.e., $P(w|a = n)$, and $\theta(m)$ to represent the multinomial distribution over topics for individual attribute m, i.e., $P(a)$. Figure 2 shows a graphical model, illustrating how to estimate the distribution over topics. In Figure 2, a refers to the assignment of word tokens to topics, and $\phi(n)$, $\theta(m)$, and a are latent variables that are required to be estimated. Box K indicates that, in each topic a, the sampling of $\phi(n)$ is repeatedly conducted until K topics are generated. Box L_m represents that, in each individual attribute, the sampling of topics/words is repeatedly conducted till L_m words are generated. Box M illustrates the sampling of a distribution over topics in each individual attribute m for a total of M individual attributes. The variables α and β, usually treated as constants, are hyper-parameters of Dirichlet prior. Consequently, the probability of the ith word in an individual attribute m is:

$$P(w_i|m) = \sum_{j=1}^{K} P(w_i|a_i = j)P(a_i = j|m), \tag{1}$$

where a_i refers to a latent variable indicating the topic from which the ith word was drawn. $P(w_i|a_i = j)$ indicates the probability of the ith word in topic j; $P(a_i = j|m)$ refers to the probability of the ith topic sampled for the ith word token for individual attribute m.

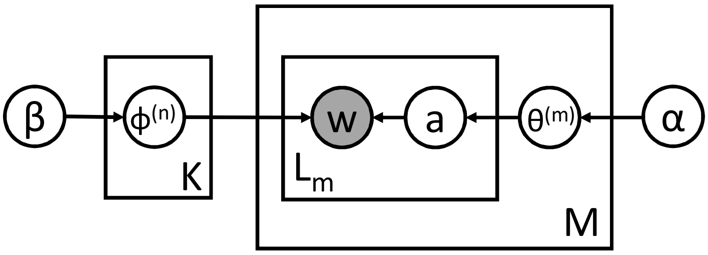


Figure 2. Graphical model for psychological feature modeling. Shaded and unshaded circles indicate observed and unobserved variables, respectively. Arrows refer to the dependences between variables. Boxes indicate repetitions of sampling steps.

Since Gibbs sampling can efficiently and easily extract topics from a big dataset [26], the two posterior (conditional) distributions over a (the assignment of word token to topics) are estimated using it. Based on that, the individual attribute (topic distribution) can be easily depicted. More details about how this technique works can be found in Griffiths and Steyvers [25].

A group of word indices w_i and individual attribute indices m_i in each word token i represent the individual attribute. Each word token is considered in turn, and the probability of assigning the current word token to each topic is estimated, conditioned on the topic assignments to all other word tokens. A topic is sampled and forms a new topic assignment for this word token based on the conditional distribution. This conditional distribution (posterior distribution) can be written as $P(a_i = n|\mathbf{a_{-i}}, w_i, m_i, \cdot)$. According to Griffiths and Steyvers [25], it can be calculated by:

$$P(a_i = n|\mathbf{a_{-i}}, w_i, m_i, \cdot) \propto \frac{C^{WL}_{w_i n} + \beta}{\sum\limits_{w=1}^{W} C^{WL}_{wn} + W\beta} \cdot \frac{C^{ML}_{m_i n} + \alpha}{\sum\limits_{l=1}^{L} C^{ML}_{m_i l} + L\alpha}, \tag{2}$$

where $a_i = n$ indicates the topic assignment of token i to topic n, a_{-i} refers to the topic assignments of all topic tokens except token i, and " $\cdot$ " refers to all other known or observed data. In addition, C^{WL} and C^{ML} are matrices of counts with $W \times L$ and $M \times L$ dimensions, respectively. C^{WL}_{wn} refers to the number of times, excluding the current instance i, word w is assigned to topic n and C^{ML}_{mn} refers to the number of times, excluding the current instance i, and topic n is assigned to some word token in individual attribute m. As a result, the left part of Equation (2) indicates the probability of word w in topic n, while the right part indicates the probability of topic n in individual attribute m. Eventually, the word-topic distribution ($\hat{\varphi}$) and semantic pattern ($\hat{\theta}$) can be estimated by:

$$\hat{\varphi}_i^{(j)} = \frac{C_{ij}^{WL} + \beta}{\sum_{k=1}^{W} C_{kj}^{WL} + W\beta}, \tag{3}$$

$$\hat{\theta}_j^{(m)} = \frac{C_{mj}^{ML} + \alpha}{\sum_{k=1}^{L} C_{mk}^{ML} + L\alpha}. \tag{4}$$

One parameter—the expected number of topics—is required for model training. A standard metric, perplexity, measuring how well a probability distribution works for a prediction—generalization performance—is used to do so [22]. Perplexity presents an exponentially decreasing trend in likelihood of training data, thus the lower the perplexity value is, the better generalization performance the model has. The perplexity is defined as:

$$Perplexity = \exp\left\{-\frac{\sum_{m=1}^{M} \log p(\mathbf{w_m}|M)}{\sum_{m=1}^{M} L_m}\right\}, \tag{5}$$

where L_m refers to the number of words in each people, $\mathbf{w_m}$ refers to the words of people m, and $p(\mathbf{w_m}|M)$ can be computed by Equation (1).

2.3. Mental Unity Measuring

According to Definition 2, we compute the similarity of topic distributions between the individuals' attributes, $P(a)$, to measure the mental unity of a crowd. Kullback–Leibler (KL) divergence is used to represent the similarity. KL divergence is a standard function to measure the difference between two topic distributions p and q in two corresponding documents [27]:

$$KL(p,q) = \sum_{i=1}^{T} p_i \log_2 \frac{p_i}{q_i}, \tag{6}$$

where T refers to the number of topics; p_i and q_i refer to the distribution of topic i in document p and q, respectively. If two documents are identical, the value of KL is zero. In other words, the value of KL is closer to zero, the two documents are more similar. In this work, p and q refer to the topic distributions of the individuals' attributes. Since KL divergence is asymmetric, a symmetric transformation of KL, KL', is used:

$$KL'(p,q) = \frac{1}{2}[KL(p,q) + KL(q,p)]. \tag{7}$$

In this case, the mental unity can be quantitatively defined as certain KL' value between involved individuals in a crowd. KL' can also be used as an indicator to show the magnitude of the mental unity, the greater the KL' is, the higher likelihood the individuals tend to be involved in the same event.

3. Case Study

The proposed approach is implemented for exploring the crowd in a protest—"Fight for $15"—occurring in more than 200 cities across the USA on 15 April 2015. This protest was claimed as the largest protest by low-wage workers in USA history and were calling for a minimum wage of

$15 per hour. The participants were mainly labor activities and fast-food workers from large chains, such as McDonald and Walmart [28].

"Fight for $15" activity began in 2012 in which two-hundred fast-food workers walked out of the job for the increase of their minimum wage and union rights. Today, this event has become a movement in over 300 cities across the world. The participants include fast-food workers, home health aides, child care teachers, airport workers, adjunct professors, retail employees and underpaid workers everywhere [29].

3.1. Data and Experiment Design

The geotagged tweets posted within the USA continent during May 2014 and August 2015 were downloaded through Twitter Streaming API [30]. The textual information of the tweets is used to infer users' psychological features and the associated geospatial information is used to further investigate the detected crowd's geospatial patterns. Twitter data has been widely used to understand human activity and mobility patterns [31,32]. Since the API only allows users to download no more than 1% tweets of the amount, only those 1% of data can be used to explore the crowd involved in the protest in this study. Despite the limitation of the API, there have been studies (e.g., Sakaki et al. [33–35]), to a certain extent, achieving the success of using it to detect events.

The experiment aims to test whether the individuals involved in the protest can be successfully detected by our proposed approach. We assume that the Twitter users who have posted any single tweet with "Fightfor15" key word and "#fightfor15" hashtag are the individuals of the crowd related to the "Fight for $15" event. Among those "Fight for $15" users, we want to investigate how many can be identified as the individuals of the crowd using our proposed approach. We use the data before the protest to detect the crowd and subsequently explore its pattern based on the data on the protest day.

A total of 152,573 geotagged tweets posted from 269 users were finally obtained after filtering noise users—(1) all their tweets are posted from an identical location and (2) all their tweets repeat the same word, hashtag and symbol "@"—as well as the users who have no tweets posted before the protest. The top-left map in Figure 3 shows how the tweets geospatially distributed during May 2014 and August 2015. It is obvious that most of tweets are located along the east coast and west coast and around Lake Michigan. Moreover, some trajectories (lines), probably representing the migration of the users, can be identified. The other two maps at the bottom in Figure 3 plot tweet density in each state before and during the protest happening. The change of the density in states may briefly illustrate the migration pattern of the participants—moving from the states with lower tweet density before the protest to the higher ones during the protest.

3.2. Psychological Feature Patterns

Topic distributions for each users are inferred to depict the individual's attribute. In order to compute the perplexity in Equation (5), the text data from each users are randomly split into two parts: 80% of them are used for model training, while the rest are used to evaluate the model. Figure 4 illustrates the distribution of the perplexity over tested topic numbers. It shows that the perplexity decreases flatly from 230 topics, thus 230 is used as the number of topics to train the topic model. Hyper-parameters of Dirichlet prior, α and β, are assigned as 0.01 and computed by $50/T = 0.5$ (T refers to the number of topics), respectively, since they are reasonable choices for topic models [26].

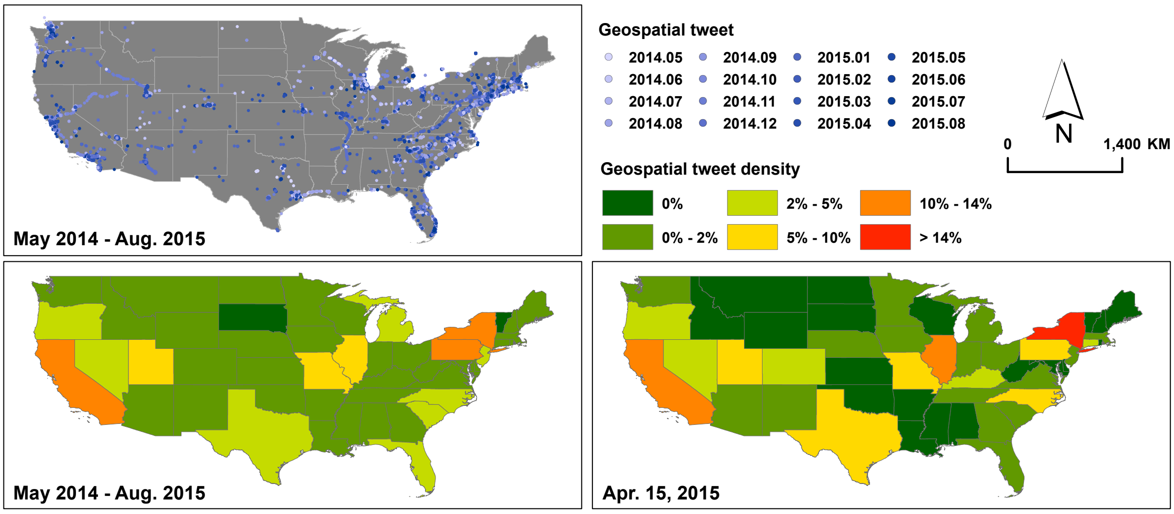


Figure 3. Maps of geospatial tweets related to the “Fight for $15” protest across the US continent.

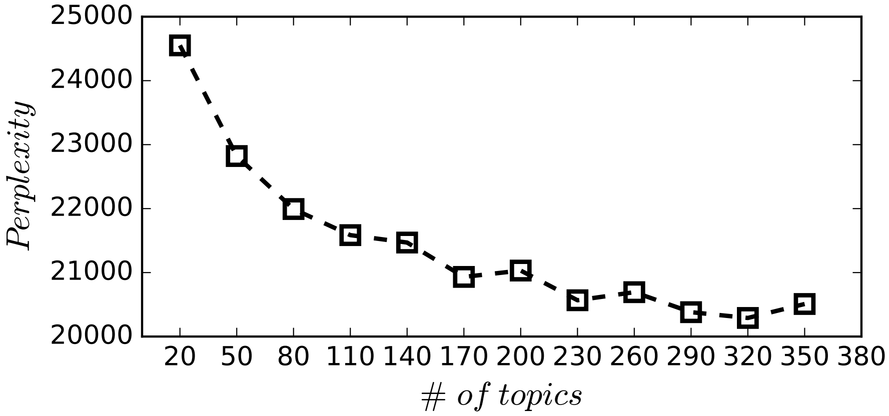


Figure 4. Perplexity distribution.

The distributions over 230 topics represent the attribute of each individual. Table 1 gives an example of the top 10 topic distributions of a randomly selected user. Each topic relates to a probability. The involved topics of each user can be various. Different words and their probability shape the semantic pattern of the topics. The meaning of each topic can be inferred based on the involved words. Among those 230 topics, we find three topics that are most relevant to the protest, as shown in Table 2, the others are various. Despite the fact that these three topics all involve some relevant key words, Topic #176 is the most relevant one since the most representative key words “#fightfor15” which is the official Twitter hashtag used for the “Fight for $15” movement and “strike fast food” only exists in this topic. In addition, the rest of the key words—“workers”, “fight for”, “wage”, “worker”, “union”, “fight” and “wages”—all hit the key content of the event according to the “Fight for $15” movement official website [29]. Consequently, we consider that the users having Topic #176 are the ones that directly participate in the protest. In this context, Table 1 can tell us that the user can be a participant of the protest; nevertheless, she/he also involves a protest-irrelevant attribute represented by topic #60 (only protest-irrelevant key words involved) with higher probability.

Table 1. An example of an individual's attribute representation.

Topic ID	$\mathbf{P(a)}$
60	0.217
176	0.152
156	0.149
198	0.148
38	0.12
92	0.112
179	0.037
190	0.029
87	0.005
84	0.002

Table 2. "Fight for $15" topic semantics.

Topic #74		Topic #176		Topic #187	
Word	$\mathbf{P(w\|a)}$	**Word**	$\mathbf{P(w\|a)}$	**Word**	$\mathbf{P(w\|a)}$
walmart	0.031	workers	0.06	wage	0.014
walmart strikers	0.027	fight for	0.038	awww	0.008
oakland	0.026	wage	0.035	brilliant	0.008
hellaodub	0.019	#fightfor15	0.021	demon	0.007
live	0.018	worker	0.021	greed	0.007
black	0.017	strike fast food	0.021	crooks and liars	0.007
our walmart	0.014	union	0.021	elected	0.007
blend	0.013	seiu	0.018	segment	0.006
hella	0.012	fight	0.014	gag	0.006
love	0.01	wages	0.013	laughable	0.005

3.3. Detected Crowd

KL' is computed for every pair of users, and the value of KL' of a total of 36,046 user pairs ranges from 0.21 to 23.25. To determine if two users are similar in terms of their attributes (topic distributions), a proper value of KL' needs to be used. We calculate mean ratio of topic intersection between two users to find this proper value. More specifically, we first group the user pairs according to their KL' value. The KL' values are split into 24 groups based on an interval of one as a result of 24 groups of user pairs. Then, the KL' value of these 24 groups ranges from (0,1] to (23,24]. Subsequently, for the user pairs in each group, ratio of the identical topics (only the topics of which $P(a)$ is none-zero count), R, is computed by the following equation:

$$R = \frac{N(T_i)}{min\{N_{u1}(T), N_{u2}(T)\}}, \tag{8}$$

where $N(T_i)$ refers to a number of identical topics of a pair of users; $N_u(T)$ refers to the number of topics of user u, u_1 and $u_2 \in$ the pair of users. Eventually, the mean ratio of each group, $\bar{R}$, is computed by:

$$\bar{R} = \frac{\sum R}{N(u)}, \tag{9}$$

where R is computed by Equation (8); $\sum R$ is the sum of R for all user pairs in a group; $N(u)$ refers to the number of user pairs in the group. The distribution of $\bar{R}$ over those 24 groups is plotted, as shown in Figure 5. It can be observed that the mean ratio dramatically decreases over the groups of which the KL' boundaries are greater than 5. In other words, the user pairs in the first five KL' groups are the most similar. In the end, the users whose KL' is no greater than 5 are identified as the involved individuals mentally unified as the crowd.

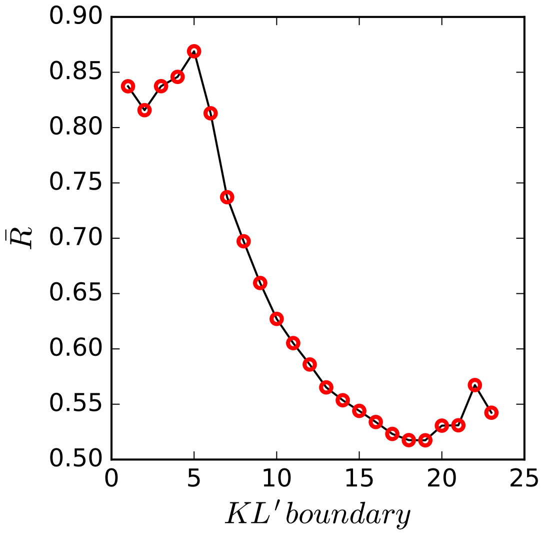
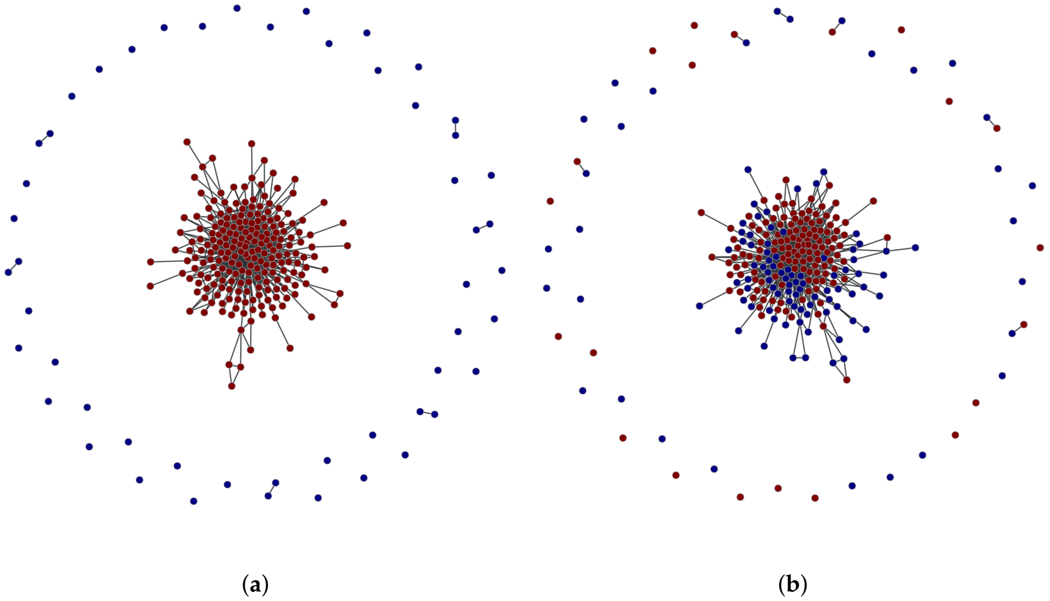


Figure 5. Mean ratio of identical topics of each KL' group.

The detected crowd consists of 218 users, represented by a network as shown in Figure 6a. The linked red points in the middle are those 218 users, while the blue points indicate the users showing lower mental unity, which cannot be identified as a part of the crowd by the model. As there are 269 users related to the "Fight for $15" protest in total, our approach is able to successfully detect 81% individuals of the crowd. In addition, we find that there are 157 users involved in Topic #176 before the protest (the red points in Figure 6b), while 112 users are found only after the protest (the blue points in Figure 6b). It may infer that: (1) there could exist some users who do not show obvious connections to the protest before but actually participate in the protest on site. These "hidden participants" would yield some unexpected impacts or increase the burden of protest management; thus, it is crucial to detect such "hidden participants"; and (2) there also could exist some users who show their interests in the event long time before the protest and also participate in the protest on site. To further investigate these two inferences, we analyzed their geospatial patterns.



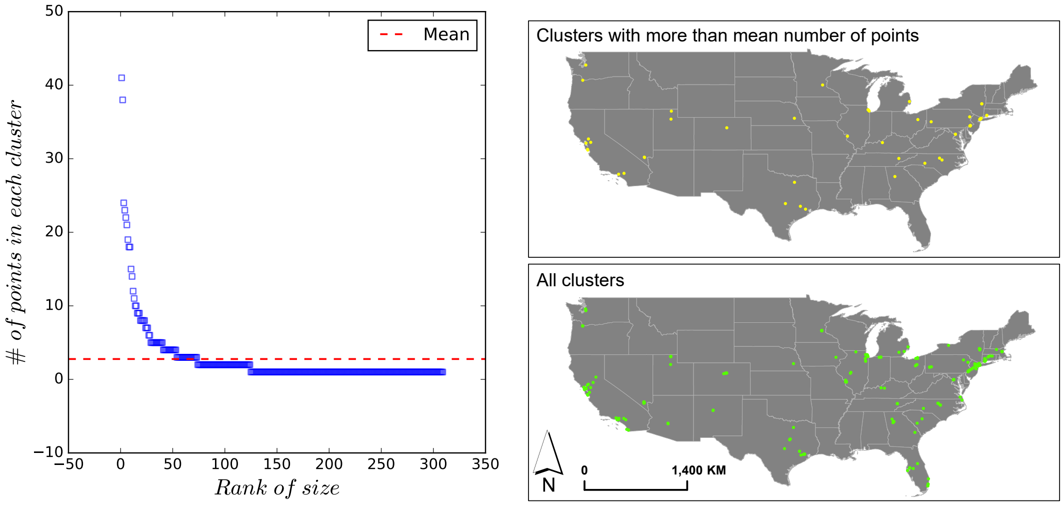
(**a**) (**b**)

Figure 6. The crowd as a network. (**a**) the detected crowd shown as linked red points; (**b**) Topic #176 related individuals shown as red points.

3.4. Geospatial Patterns

In order to infer the locations where the protests occurred on 15 April 2015, we develop spatial clustering based on DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [36] to group the tweets posted on that day. DBSCAN does not require the number of clusters in advance, which meets our demand in this study. Furthermore, 200 m is used as the distance threshold for DBSCAN since the street block is designed between 50 to 200 m in the USA, and we assume that the participants of a protest should be distributed around a street block. The density threshold of DBSCAN is assigned as one because the tweets used are only 1% of the amount tweets, and the low density enables detecting the possible locations as much as possible. Finally, 309 spatial clusters are generated to represent the locations of the "Fight for $15" protests occurring on 15 April 2015.

The number of points in each cluster ranges from 1 to 41. Figure 7a illustrates that a part of clusters (53 clusters) contains a very small number of points—the ones below the mean (three points). These clusters could be noise that some users post tweets from some other places rather than the sites where the protests occur. However, it could be that the protests are also held there, but less participants are involved, resulting in a low volume of tweets generated. Then, it makes sense that such a small number of tweets shape those "small" clusters. In this paper, we analyze two scenarios—all clusters and the clusters with more than a mean number of points–to explore the geospatial patterns of the crowd. The geospatial distributions of these two groups of clusters are plotted in Figure 7b.



(**a**) (**b**)

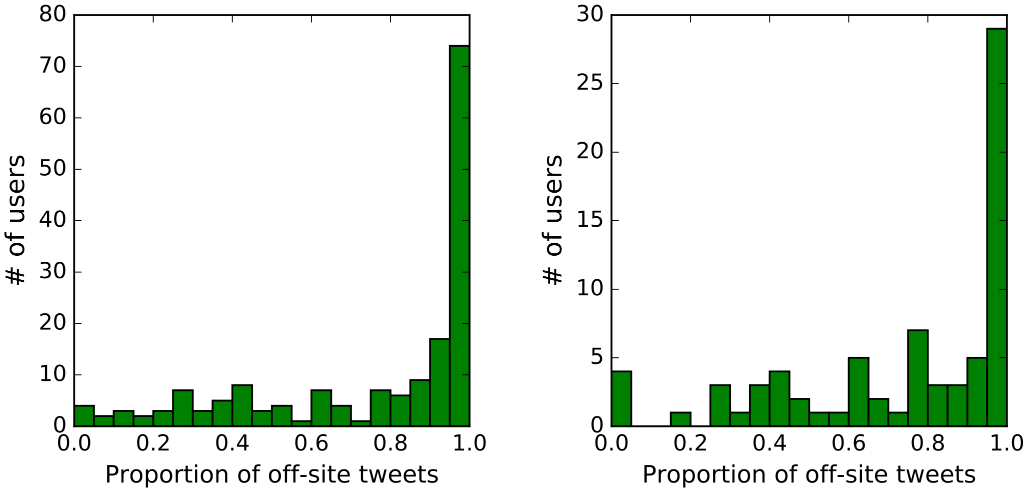
Figure 7. Distributions of spatial clusters. (**a**) distribution of number of points over clusters; (**b**) geospatial distribution.

The centroid of the clusters is used as the location of the protest, based on which a 200 m buffer is created to detect onsite users. Table 3 lists some statistics on the geospatial characteristics for the two scenarios. With regards to all clusters, 170 users are detected from 309 event spots on event day. Among them, 51 users do not tweet at the event spots before the protest, while 119 users tweet from the event spots before the protest. The ratio of user amount between these two types of users is 51/119 = 0.43. With regards to the clusters with more than a mean number of points (selected clusters), 75 users are detected from 53 event sports on event day. Among them, 18 users do not tweet at the event spots before the protest, whilst 57 users tweet from the event spots before the protest. The ratio of user amount between these two types of users is 18/57 = 0.31.

Table 3. Geospatial patterns.

	All Clusters	Selected Clusters
# of clusters	309	53
# of users	170	75
Users posting off-site tweets before event (a)	51	18
Users posting on-site tweets before event (b)	119	57
a:b	0.43	0.31

Figure 8 shows the number of users in a different proportion of the tweets posted from other spots rather than the protest spots. Apart from the users who do not tweet at the event spots before the protest, there are also some users who only post tweets from the event spots before the protests (those users whose proportion of off-site tweets is 0 in Figure 8). These users can be seen as the people working there (e.g., McDonald's shop). They complain and tweet when they are working and also participate in the protests held at their work places. Such users are the directly event-related people, which can be identified before the protest. In contrast, those users who do not tweet on the protest spots at all before the protest but suddenly participate the protest can be seen as latently event-related people, they cannot be directly observed before the event.



(**a**) (**b**)

Figure 8. Histogram of off-site twitting users. (**a**) all clusters; (**b**) the clusters with more than a mean number of points.

3.5. Discussion

Although only the 1% of tweets are used to test the model, the experiment shows that the proposed approach can detect the individuals associated with the "Fight for $15" crowd at 81% accuracy. The performance of the model would be increased along with more data being involved. The three relevant topics are inferred from this 1% of data, which may tell us that this small data sample is suitable for such large-scale crowds and collective activities research. Regarding the spatial clusters selection—using all clusters or selected clusters. Nevertheless, the geospatial patterns shown in Table 3 look different, and the values of a:b for these two groups of clusters are close. It can prove our assumption that some protests are also held at the spots where the selected clusters are located. The reason that these clusters show very low volume of tweets is most likely due to small data samples used in this case study.

The idea of detecting the crowd is to find the individuals who have similar psychological features. Individuals' attributes representing their "thoughts" are used to model the psychological features. The individuals would take the same action, participating in protests in our case study, regardless of where they are located. Such action may cause a big event, which is likely to create violence later on. Unfortunately, some of those latent individuals cannot be sensed until they take the action. The approach developed in this study uses a computational way to enable those latent individuals to be surfaced to a certain extent, which lays a foundation to some applications, such as social events prediction, human activity prediction, monitoring violent offenders, and emergency response.

However, there are still some limitations in this research: (1) we only consider what people tweet to model the individuals' attributes due to data accessibility—other social media data during the event period were unfortunately not collected, and a more comprehensive model should be developed if other types of data are involved. For example, image topic analytics has to be developed if relevant image data collected from both social media (e.g., Flickr) and/or news report are involved, and an upgraded model that is capable of dealing with the information derived from the images needs to be designed to depict the psychological feature; (2) the two thresholds of DBSCAN are assigned based on the geographical characteristics of the study area, and they may be different for other study areas. Moreover, other than DBSCAN, other clustering algorithms can be used; (3) indeed, not everyone posts relevant content before or during the event, which means that those individuals who are not able to be observed online cannot be revealed by any means. On the other hand, there is a part of users who post on social media from time to time but share their "thoughts" infrequently, which may limit the detecting accuracy.

4. Conclusions

People distributed at different locations in urban environments can be driven to gather at certain places by certain events, which may result in both positive and negative effects on urban systems and the society. Thus, it is necessary to detect the involved crowd in order to analyze the possible influence. This paper presents an approach to model the crowd based on the concept of psychological crowd. Topic modeling was used to infer the individuals' attributes representing the psychological feature of the crowd. Then, the similarity of the attribute between individuals was computed to measure the mental unity of the crowd. Eventually, a psychological crowd was shaped by the individuals showing similar mental unity.

The approach was tested for a national event—"Fight for $15" protest–occurring in the USA in April, 2015. Twitter data was used as a proxy of the event. The experiment shows that our approach was able to successfully detect 81% of individuals in the crowd, where not only the observed users before the event were detected but also those unobserved users were also uncovered by analyzing their geospatial patterns. The proposed approach can effectively represent the psychological feature and measure the mental unity of the crowd, which can be implemented for uncovering human activity patterns from motivation level.

We believe the capability of the detection of the latent social event users may benefit a set of urban management, ranging from policy making to emergency response, and provide a new tool for the field of psychology to study the psychological behavior of large-scale crowds and for the GIS scientists and social scientists to better understand the collective human behavior before and during the occurrence of big social events. In the future, a broader range of datasets will be coupled to improve the detecting accuracy and a more comprehensive model will be developed to draw the psychological features.

Acknowledgments: We would like to thank Trinn Christoph for his discussion for the original idea of this research. This research was partly supported by the funding from Klaus Tschira Foundation, Heidelberg. We acknowledge the financial support of the Deutsche Forschungsgemeinschaft and Ruprecht-Karls-Universität Heidelberg within the funding programme Open Access Publishing.

Author Contributions: Wei Huang, Hongchao Fan and Alexander Zipf designed the experiments; Wei Huang performed the experiments; Wei Huang and Hongchao Fan analyzed the data; Wei Huang wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nye, R.A. *The Origins of Crowd Psychology: Gustave LeBon and the Crisis of Mass Democracy in the Third Republic*; Sage Publications: Beverly Hills, CA, USA; London, UK, 1975; Volume 2.
2. Le Bon, G. *The Crowd: A Study of the Popular Mind*; Macmillan: New York, NY, USA, 1921.
3. Turner, R.H.; Killian, L.M. *Collective Behavior*, 3rd ed.; Prentice-Hall: Englewood Cliffs, NJ, USA, 1987.
4. Helbing, D.; Molnár, P.; Farkas, I.J.; Bolay, K. Self-organizing pedestrian movement. *Environ. Plan. B Plan. Des.* **2001**, *28*, 361–383.
5. Moussaïd, M.; Helbing, D.; Theraulaz, G. How simple rules determine pedestrian behavior and crowd disasters. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 6884–6888.
6. Sieben, A.; Schumann, J.; Seyfried, A. Collective phenomena in crowds—Where pedestrian dynamics need social psychology. *PLoS ONE* **2017**, *12*, e0177328.
7. Von Krüchten, C.; Schadschneider, A. Empirical study on social groups in pedestrian evacuation dynamics. *Phys. A Stat. Mech. Appl.* **2017**, *475*, 129–141.
8. Templeton, A.; Drury, J.; Philippides, A. From mindless masses to small groups: Conceptualizing collective behavior in crowd modeling. *Rev. Gen. Psychol.* **2015**, *19*, 215.
9. Sun, L.; Axhausen, K.W.; Lee, D.H.; Huang, X. Understanding metropolitan patterns of daily encounters. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 13774–13779.
10. Järv, O.; Ahas, R.; Witlox, F. Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transp. Res. Part C Emerg. Technol.* **2014**, *38*, 122–135.
11. Alexander, L.; Jiang, S.; Murga, M.; González, M.C. Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transp. Res. Part C Emerg. Technol.* **2015**, *58*, 240–250.
12. Ferrara, E.; De Meo, P.; Catanese, S.; Fiumara, G. Detecting criminal organizations in mobile phone networks. *Expert Syst. Appl.* **2014**, *41*, 5733–5750.
13. Fast, S.M.; González, M.C.; Wilson, J.M.; Markuzon, N. Modelling the propagation of social response during a disease outbreak. *J. R. Soc. Interface* **2015**, *12*, doi:10.1098/rsif.2014.1105.
14. Wang, P.; González, M.C.; Hidalgo, C.A.; Barabási, A.L. Understanding the spreading patterns of mobile phone viruses. *Science* **2009**, *324*, 1071–1076.
15. González, M.C.; Hidalgo, C.A.; Barabasi, A.L. Understanding individual human mobility patterns. *Nature* **2008**, doi:10.1038/nature06958.
16. Song, C.; Qu, Z.; Blumm, N.; Barabási, A.L. Limits of predictability in human mobility. *Science* **2010**, *327*, 1018–1021.
17. Lima, A.; Stanojevic, R.; Papagiannaki, D.; Rodriguez, P.; González, M.C. Understanding individual routing behaviour. *J. R. Soc. Interface* **2016**, *13*, doi:10.1098/rsif.2016.0021.
18. Jiang, S.; Ferreira, J.; González, M.C. Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *IEEE Trans. Big Data* **2017**, *3*, 208–219.
19. Chaniotakis, E.; Antoniou, C.; Aifadopoulou, G.; Dimitriou, L. Inferring activities from social media data. In Proceedings of the 96th Transportation Research Board Annual Meeting, Washington, DC, USA, 8–12 January 2017.
20. Guellil, I.; Boukhalfa, K. Social big data mining: A survey focused on opinion mining and sentiments analysis. In Proceedings of the 12th IEEE International Symposium on Programming and Systems (ISPS), Algiers, Algeria, 28–30 April 2015; pp. 1–10.
21. Liao, L.; Fox, D.; Kautz, H. Extracting places and activities from gps traces using hierarchical conditional random fields. *Int. J. Robot. Res.* **2007**, *26*, 119–134.
22. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
23. Griffiths, T.; Steyvers, M. A probabilistic approach to semantic representation. In Proceedings of the 24th Annual Conference of the Cognitive Science Society, Fairfax, VA, USA, 8–10 August 2002; pp. 381–386.
24. Griffiths, T.; Steyvers, M. Prediction and semantic association. In *Advances in Neural Information Processing Systems 15*; MIT Press: Cambridge, MA, USA, 2002; pp. 11–18.
25. Griffiths, T.L.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5228–5235.

26. Steyvers, M.; Griffiths, T. Probabilistic topic models. In *Handbook of Latent Semantic Analysis*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2007; Volume 427, pp. 424–440.
27. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151.
28. Fight for $15: Workers Across US Protest to Raise Minimum Wage–As It Happened. Available online: https://www.theguardian.com/us-news/2015/apr/15/fight-for-15-minimum-wage-protests-new-york-los-angeles-atlanta-boston (accessed on 15 April 2015).
29. Fight for $15 Organization. Available online: https://fightfor15.org/about-us/ (accessed on 29 November 2012).
30. Twitter Developer Documents. Available online: https://developer.twitter.com/en/docs (accessed on 1 January 2017).
31. Jurdak, R.; Zhao, K.; Liu, J.; AbouJaoude, M.; Cameron, M.; Newth, D. Understanding human mobility from Twitter. *PLoS ONE* **2015**, *10*, e0131469.
32. Huang, W.; Li, S. Understanding human activity patterns based on space-time-semantics. *ISPRS J. Photogramm. Remote Sens.* **2016**, *121*, 1–10.
33. Atefeh, F.; Khreich, W. A survey of techniques for event detection in twitter. *Comput. Intell.* **2015**, *31*, 132–164.
34. Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake shakes Twitter users: Real-time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; ACM: New York, NY, USA, 2010; pp. 851–860.
35. Taxidou, I.; Fischer, P.M. Online analysis of information diffusion in twitter. In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, 7–11 April 2014; ACM: New York, NY, USA, 2014; pp. 1313–1318.
36. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996; Volume 96, pp. 226–231.